Эконометрика — Совбак ВШЭ и РЭШ, 2021 midterm

Совбак ВШЭ и РЭШ | Эконометрика | 2021 | midterm

QUESTION 1

Part 1 — 25 points

Suppose (Y_i, X_{1i}, X_{2i}) satisfy the four least-squares assumptions for causal inference in the multiple-regression model discussed in class. In addition,

$$\mathrm{Var}(u_i \mid X_{1i}, X_{2i}) = 4, \qquad \mathrm{Var}(X_{1i}) = 6.$$

A random sample of size $n = 400$ is drawn from the population.

(a) (9 points) Assume that X_1 and X_2 are uncorrelated. Compute the asymptotic unconditional variance of $\widehat{\beta}_1$.

(b) (9 points) Assume that

$$\mathrm{corr}(X_1, X_2) = 0.5.$$

Compute the asymptotic unconditional variance of $\widehat{\beta}_1$.

(c) (7 points) Comment on the following statements:

When X_1 and X_2 are correlated, the variance of $\widehat{\beta}_1$ is larger than it would be if X_1 and X_2 were uncorrelated. Thus, if you are interested in β_1, it is best to leave X_2 out of the regression if it is correlated with X_1.

QUESTION 2

Part 2 — 30 points

Table 2 reports regressions estimated using data on employees in a developing country. The data set contains information on more than 10,000 full-time, full-year workers. Each worker's highest educational qualification is either a high-school diploma or a bachelor's degree. Ages range from 25 to 40. The data also contain region, gender, and age.

Table 1. Variable definitions

Name	Description
AWE	Logarithm of average weekly earnings, in 2007 units
High School	Binary variable equal to 1 for a high-school graduate and 0 for a college gradua
Male	Binary variable equal to 1 for male and 0 for female
Age	Age in years
North	1 if region is North, 0 otherwise
East	1 if region is East, 0 otherwise
South	1 if region is South, 0 otherwise
West	1 if region is West, 0 otherwise

Table 2. Regressions of log average weekly earnings

Standard errors are in parentheses.

Regressor	(1)	(2)	
High-school graduate, X_1	0.352 (0.021)	0.373 (0.021)	0.371
Male, X_2	0.458 (0.021)	0.457 (0.020)	0.451
Age, X_3	—	0.011 (0.001)	0.011
North, X_4	—	—	0.175
South, X_5	—	—	0.103

Regressor	(1)	(2)	
East, X_7	—	—	-0.102
Intercept	12.840 (0.018)	12.471 (0.049)	12.390
F statistic for zero regional effects	—	—	
SER	1.026	1.023	
R^2	0.0710	0.0761	
n	10,973	10,973	

(a) (2 points) For every coefficient in all three regressions, add * for significance at the 5% level and ** for significance at the 1% level.

Using regression (1):

(b) (3 points) Is the estimated high-school earnings difference statistically significant at the 5% level? Construct a 95% confidence interval.

(c) (3 points) Is the estimated male-female earnings difference statistically significant at the 5% level? Construct a 95% confidence interval.

Using regression (2):

(d) (3 points) Is age an important determinant of earnings? Use an appropriate test and/or confidence interval.

(e) (3 points) Alvo is a 30-year-old male college graduate, and Kal is a 40-year-old male college graduate. Construct a 95% confidence interval for the expected difference between their earnings.

Using regression (3):

(f) (3 points) Are there important regional differences? Use an appropriate hypothesis test.

(g) Juan is a 32-year-old male high-school graduate from the North. Mel is a 32-year-old male college graduate from the North. Ari is a 32-year-old male college graduate from the East.

1. **(3 points)** Construct a 95% confidence interval for the difference in expected earnings between Juan and Mel.

2. **(4 points)** Explain how you would construct a 95% confidence interval for the difference in expected earnings between Juan and Ari.

(h) (3 points) Regression (2) was re-estimated using 5,000 randomly selected observations from 1993, with earnings converted into 2007 units using the Consumer Price Index:

$$\widehat{\log AWE} = 9.32 + 0.301\, HighSchool + 0.562\, Male + 0.011\, Age,$$

with standard errors

$$(0.20) \qquad (0.019) \qquad (0.047) \qquad (0.002),$$

and

$$SER = 1.25, \qquad \bar{R}^2 = 0.08.$$

Compared with the 2007 regression in column (2), was there a statistically significant change in the coefficient on High School?

(i) (3 points) In all regressions above, the coefficient on High School is positive, large, and statistically significant. Does this provide strong evidence of high returns to schooling in the labor market? Explain.

QUESTION 3

Part 3 — 20 points

A researcher collects data on houses sold in one neighborhood during the past year and obtains the following regressions.

Table 3. Regressions of house price on house characteristics

Dependent variable: $\ln(Price)$. Standard errors are in parentheses.

Regressor	(1)	(2)	(3)	
Size	0.00042 (0.000038)	—	—	
$\ln(Size)$	—	0.69 (0.054)	0.68 (0.087)	0.57
$[\ln(Size)]^2$	—	—	—	0.0078
Bedrooms	—	—	0.0036 (0.037)	
Pool	0.082 (0.032)	0.071 (0.034)	0.071 (0.034)	0.071<
View	0.037 (0.029)	0.027 (0.028)	0.026 (0.026)	0.027<
Pool × View	—	—	—	
Condition	0.13 (0.045)	0.12 (0.035)	0.12 (0.035)	0.12<
Intercept	10.97 (0.069)	6.60 (0.39)	6.63 (0.53)	7.02
SER	0.1026	1.023	—	
R^2	0.0710	0.0761	—	

Definitions: $Price$ is sale price in dollars; $Size$ is square feet; Bedrooms is number of bedrooms; Pool is 1 if the house has a swimming pool; View is 1 if the house has a nice view; Condition is 1 if the real-estate agent reports that the house is in excellent condition.

(a) (3 points) Using column (1), what is the expected change in price from building a 1,500-square-foot addition? Construct a 99% confidence interval for the percentage change in price.

(b) (3 points) How is the coefficient on $\ln(Size)$ interpreted in column (2)? What is the effect of doubling house size on price?

(c) (3 points) Using column (2), estimate the effect of a view on price. Construct a 99% confidence interval. Is the effect statistically different from zero?

(d) (3 points) Using column (3), calculate the effect of adding two bedrooms. Is the effect statistically significant? Which variable, size or number of bedrooms, appears more important for determining price?

(e) (3 points) Is the coefficient on Condition significant in column (4)?

(f) (5 points) Is the interaction between Pool and View significant in column (5)? Find the effect of adding a view for a house with a pool and for a house without a pool.

QUESTION 4

Part 4 — 25 points

Demand for a commodity is

$$Q = \beta_0 + \beta_1 P + u,$$

where Q is log quantity, P is log price, and u collects other demand determinants.

Supply is

$$Q = \gamma_0 + \gamma_1 P + v,$$

where v collects other supply determinants. Assume

$$E(u) = E(v) = 0, \qquad \mathrm{Var}(u) = \sigma_u^2, \qquad \mathrm{Var}(v) = \sigma_v^2, \qquad \mathrm{Cov}(u, v) = 0.$$

(a) (5 points) Solve the simultaneous equations to show how Q and P depend on u and v.

(b) (5 points) Derive the means of P and Q.

(c) (5 points) Derive $\mathrm{Var}(P)$, $\mathrm{Var}(Q)$, and $\mathrm{Cov}(Q, P)$.

(d) A large random sample of (Q_i, P_i) is collected, and Q_i is regressed on P_i.

1. **(5 points)** Use your answers to parts (b) and (c) to derive the population regression coefficients.

2. **(5 points)** A researcher treats the slope as an estimate of the demand slope β_1. Is the estimated slope too large or too small?